# The Weaknesses of Automated Taxonomy: A Case Study

Steffanie Cain, Yohta Shimizu, Erika Weir
School of Information Sciences, University of Illinois at Urbana-Champaign; in partnership with the University of Illinois Press

**ILLINOIS**
School of Information Sciences

## Definition

The term automated taxonomy, for our purposes, describes a system used to analyze text for the purpose of creating a list of subject or topical headings for item categorization that features limited or no human input into selection.

## Introduction

As access to the internet continues to increase, our information sources are shifting to online and digital spheres at an exponential rate. However, the extraordinary amount of digital information poses a findability challenge to those seeking information online. To improve item findability for users, more platforms are incorporating the use of subject headings, tags, and other forms of descriptive metadata. Although there are many approaches to assigning subject headings, most fall within the following categories: the automated approach and human approach.

| | Pros | Cons | Typical Use Cases |
|---|---|---|---|
| **Automated Taxonomy** | Speed Capacity | Inaccuracy Initial cost Maintenance cost | A need for speed in indexing, such as for time-critical information. (e.g.Newsmedia) |
| **Human Taxonomy** | Accuracy Flexibility | Labor-intensity Sustainability | Limited quantitiy of documents that require relatively high accuracy(e.g. Conventional government/library documents) |

Automated subject assignment has the potential to greatly reduce the time and effort spent on taxonomic and related activities, which is of particular benefit when attempting to categorize large bodies of work, but there are many pitfalls that can be difficult to avoid when attempting to use technology to complete such a highly subjective task.

Human taxonomic activities are inherently quite time-consuming, but provide a higher degree of discursive nuance than automated systems. This problem is magnified in the case of large platforms which host academic publications in a variety of disciplines, because of the automated system's inability to discern between disciplines.
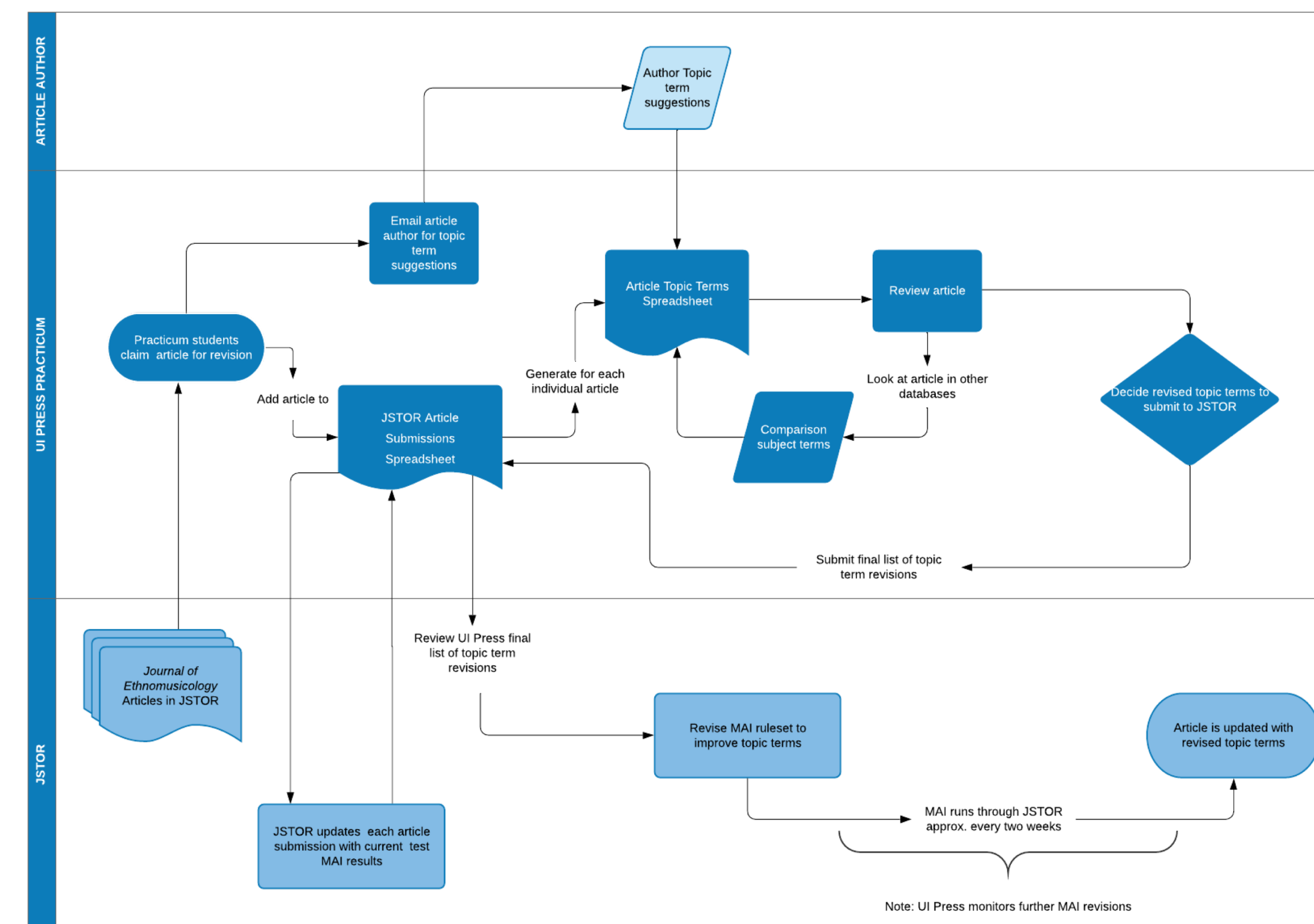
## Objective

As iSchool students working for the University of Illinois Press (herein referred to as "the Press"), we initially took an iterative approach to improve the accuracy of topics assigned to Ethnomusicology, one of the most important journals published by the Press, by manually reviewing the articles and topic lists for errors and suggesting improvements. This process led to some overall improvement of the platform's rule-based algorithms, but is not sustainable due to the labor-intense and time-consuming nature of the work.

Our long-term objective is to raise awareness of the issue based on our experiences and consequently involve more stakeholders in a constructive discussion within the field. Although we are aware of the continuous trial and error processes in the context of automation technology, we hope that our work will contribute to an increased understanding of the importance of human maintenance and input when developing automated taxonomies.
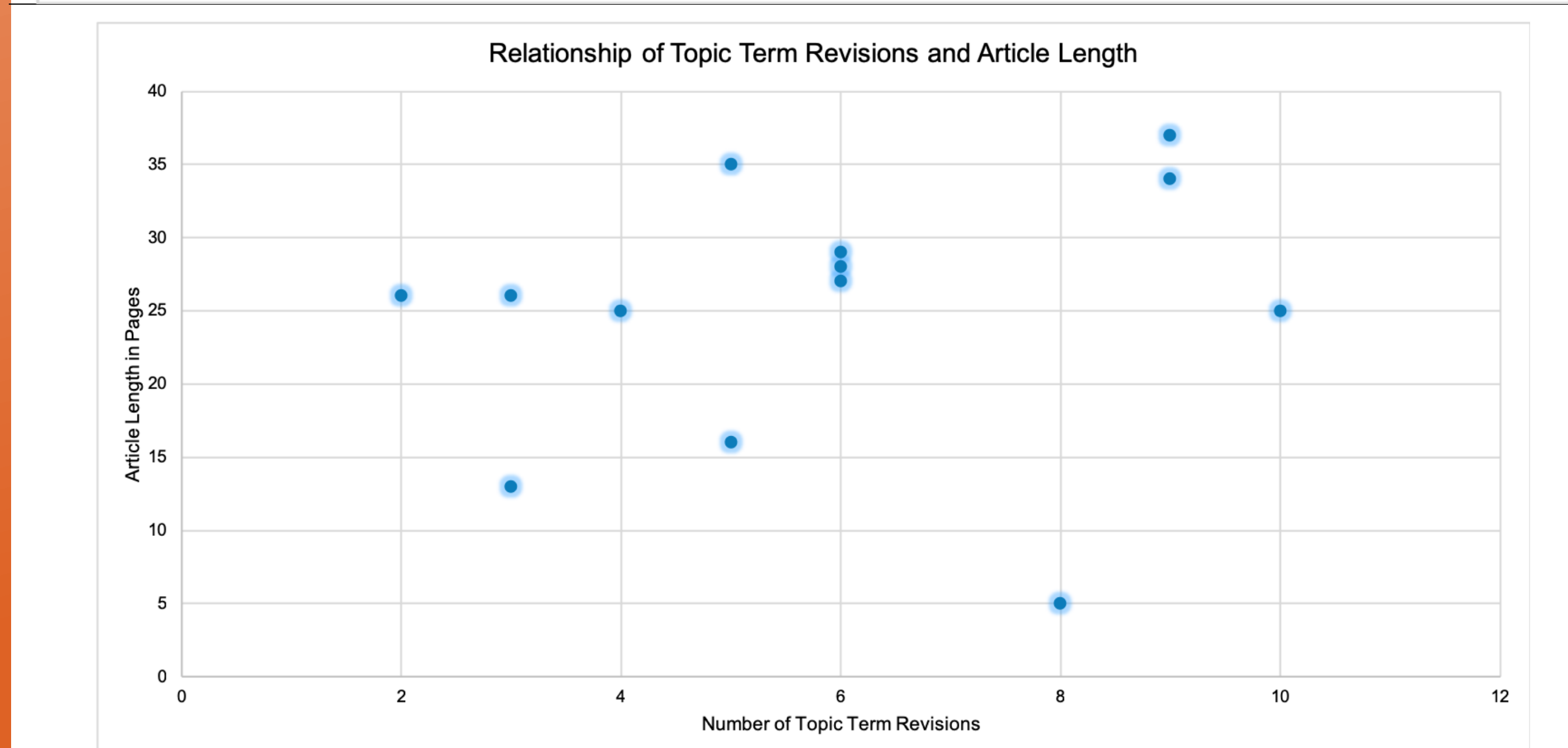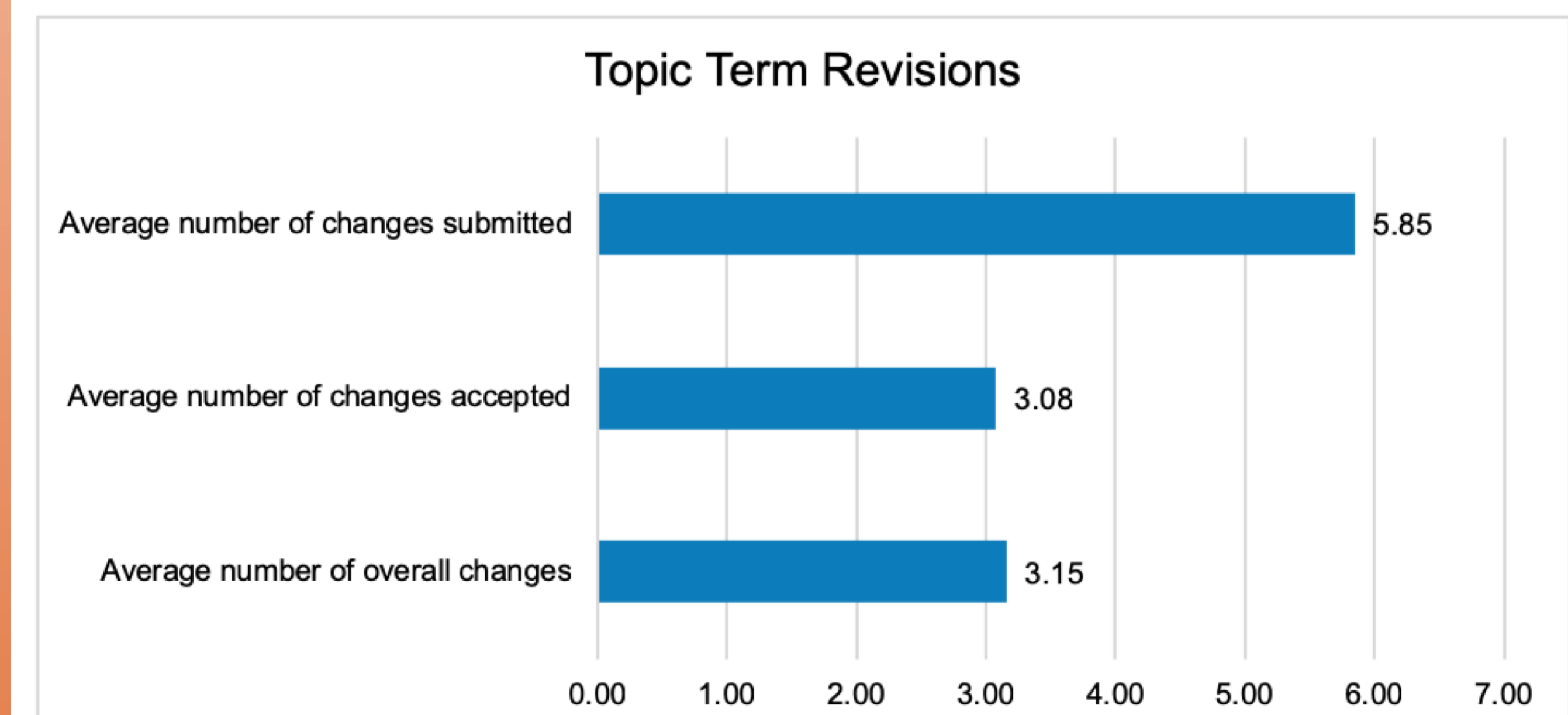
## Method

To begin the revision process, a review is conducted of the existing article topic list, the full text of the article, and, when available, comparisons are made with a list of topics suggested by the author and the subject headings assigned by other content indexers. The topic frequency and relevant topic rules are also examined to provide more information about how the initial terms are generated by the automated taxonomy system. Then, informed by the information gathered and our own analysis of the content, we suggest a new list of topics that more accurately suit the article. JSTOR reviews the suggestions and implements changes as appropriate. The changes JSTOR makes may be incomplete due to limitations on how rules can be changed without negatively affecting the rest of the corpus.



## Examples

### "Paul Simon's Graceland, South Africa, and the Mediation of Musical Meaning" by Louise Meintjes

| Original Topic List | Our Topic Suggestions | Final Topic List |
|---|---|---|
| African music Collaboration Musicians Music criticism Popular music Musical expression Music industry Music concerts | African music Collaboration Popular music Music industry Cultural appropriation Apartheid Nationalism Music semiology Globalization Music criticism | Collaboration African music Popular music African culture Black music Music semiology Apartheid Cultural appropriation Nationalism Song lyrics |

For this article, we felt that the topic Musicians was too broad to be useful and that Musical expression and Music concerts, while present in the text, were not truly representative of the article's subject matter.

We suggested the addition of the topics Apartheid, Globalization, and Nationalism which were prevalent terms in the text, and the topics Cultural appropriation and Music semiology which are contextual elements of the article.

### "'Our Angel of Salvation': Toward an Understanding of Iranian Cyberspace as an Alternative Sphere of Musical Sociality" by Laudan Nooshin

| Original Topic List | Our Topic Suggestions | Final Topic List |
|---|---|---|
| Internet Musicians Music Popular music. Musical performance Cyberspace Music festivals Musical register Vocal music Websites | Cyberspace Musical performance Popular music Public sphere Censorship Gender roles Islamic culture Women Online media Private life | Online media Public sphere Private life Popular music Musicians Cyberspace Music festivals Musical register Vocal music Websites |

We found that some topics for this article were being applied erroneously due to how the phrasing of the rule was interacting with certain terms, for example Musical register was being triggered by words like "voice" and "singing" despite not actually being present in the text. Many topics about music and the internet were judged to be too broad or not as accurate as other potential topics.

We suggested adding the topics Islamic culture, Public sphere, Censorship, Online media, Gender roles, Women, and Private life to focus more on the article's discussion of culturally-acceptable behavior for female musicians in public in Iran.

## Initial Findings

Based on our review of thirteen of the most highly used Ethnomusicology articles in JSTOR, we found the taxonomy-generated topic terms to have an accuracy rate of approximately 74.6%. On average, we submitted 5.85 topic term revisions per article; of the changes submitted, we had an average acceptance rate of 50.92%. Often, however, additional changes to inaccurate AI rule sets would result in larger overall change than the number of suggested changes adopted. Overall, our revisions resulted in average change of improved accuracy by 31.54%. Furthermore, initial findings led us to believe that there may be a relationship between the article length and overall accuracy of topic terms. However, our initial data set did not prove any correlation between the two values.

**Topic Term Revisions**

| | |
|---|---|
| Average number of changes submitted | 5.85 |
| Average number of changes accepted | 3.08 |
| Average number of overall changes | 3.15 |



Relationship of Topic Term Revisions and Article Length

## Conclusions

Use of automated text analysis tools, like taxonomies, topic modeling, and text-mining are becoming progressively more important within the field of scholarly communications. As content providers increasingly rely on the use of AI and text mining to build findability tools, information professionals need to be aware of the potential implications for our field. No algorithm or model is perfect without timely feedback and fine-tuning based on data. As O'Neil argued (2017), algorithmic models without tuning can be self-reinforcing, or misleading at best. (O'Neil, 2017, p.82). Needless to say, automated tools can reduce a significant amount of time and effort spent if used appropriately. Our work is to highlight that people should be aware of the shortcomings of the field, and work constructively for better use of the technology.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

**ILLINOIS**